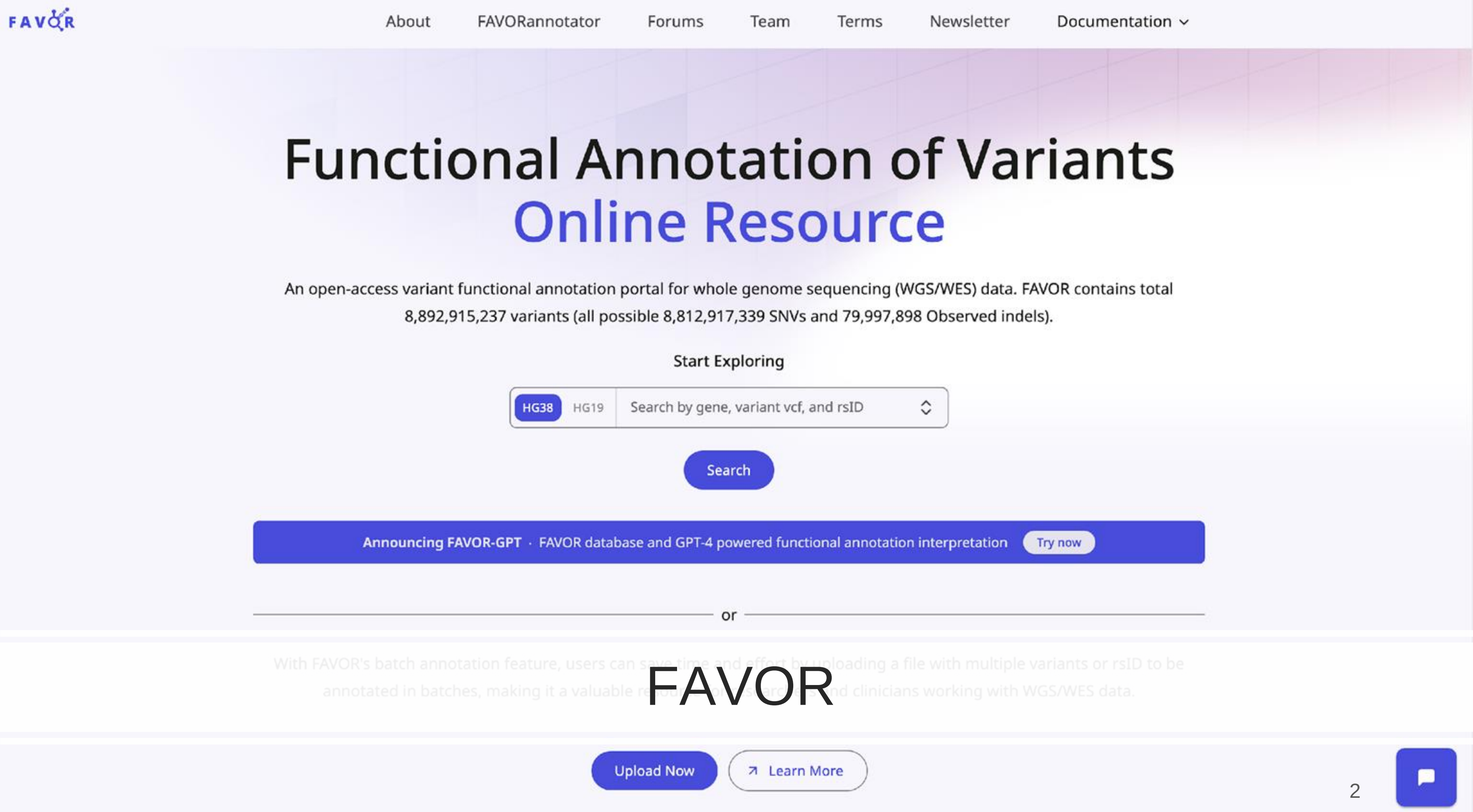
Integrating Biomedical Data Database Development with LLMs

Presenter: Hufeng Zhou

FAVOR



Functional Annotation of Variants Online Resource

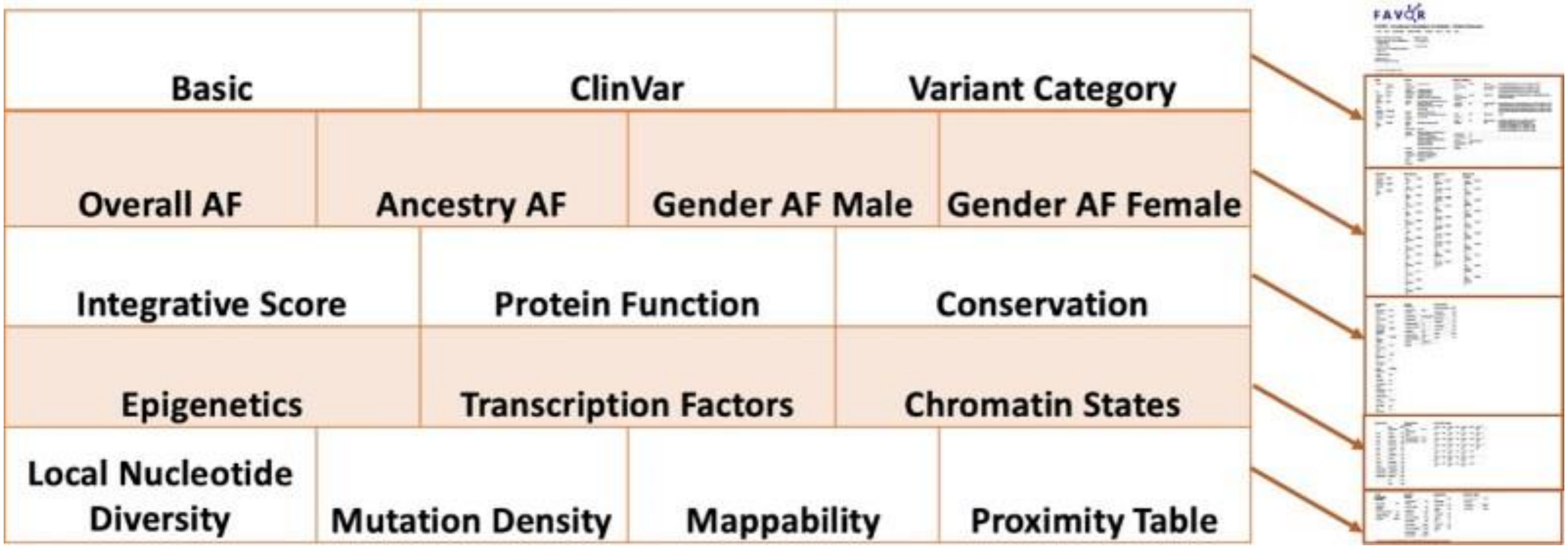
An open-access variant functional annotation portal for whole genome sequencing (WGS/WES) data. FAVOR contains total 8,892,915,237 variants (all possible 8,812,917,339 SNVs and 79,997,898 Observed indels).

Announcing FAVOR-GPT · FAVOR database and GPT-4 powered functional annotation interpretation

With FAVOR's batch annotation feature, users can save time and effort by uploading a file with multiple variants or rsID to be annotated in batches, making it a valuable resource for researchers and clinicians working with WGS/WES data.

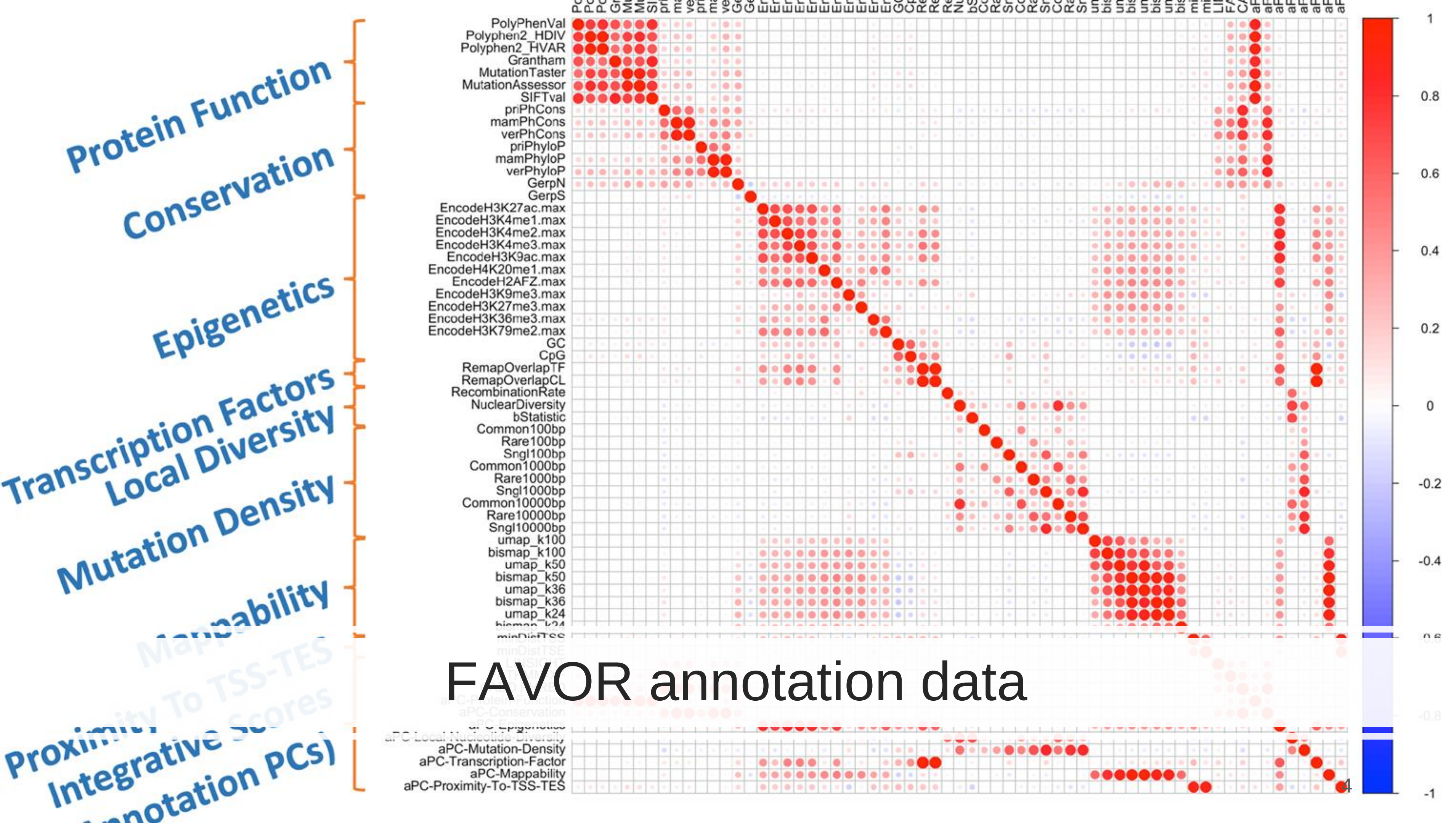
FAVOR

- Functional annotation of variants online resource and annotator for variation across the human genome

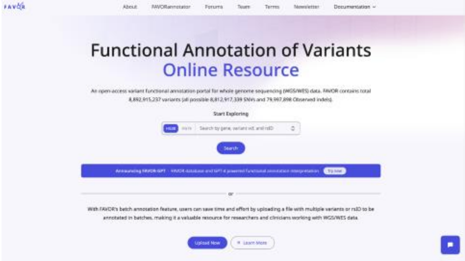


Basic
ClinVar
Variant Category
Overall AF
Ancestry AF
Gender AF Male
Gender AF Female
Integrative Score
Protein Function
Conservation
Epigenetics
Transcription Factors
Chromatin States
Local Nucleotide Diversity
Mutation Density
Mappability
Proximity Table

FAVOR annotation data



FAVOR API



Functional Annotation of Variants
Online Resource

An open-access variant functional annotation portal for whole genome sequencing (WGS/WES) data. FAVOR contains total 8,892,915,237 variants (all possible 8,812,917,339 SNVs and 79,997,898 Observed indels).

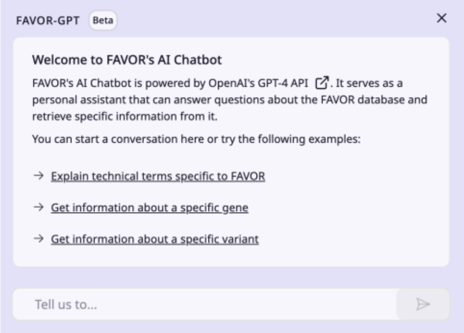
Start Exploring

Search by gene, variant vcf, and rsID

or

With FAVOR's batch annotation feature, users can save time and effort by uploading a file with multiple variants or rsID to be annotated in batches, making it a valuable resource for researchers and clinicians working with WGS/WES data.

FAVOR Web UI



Welcome to FAVOR's AI Chatbot

FAVOR's AI Chatbot is powered by OpenAI's GPT-4 API. It serves as a personal assistant that can answer questions about the FAVOR database and retrieve specific information from it.

You can start a conversation here or try the following examples:

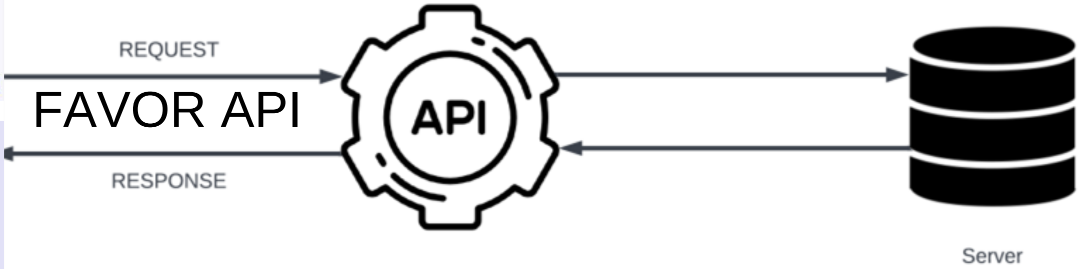
→ Explain technical terms specific to FAVOR

→ Get information about a specific gene

→ Get information about a specific variant

Tell us to...

FAVOR DB

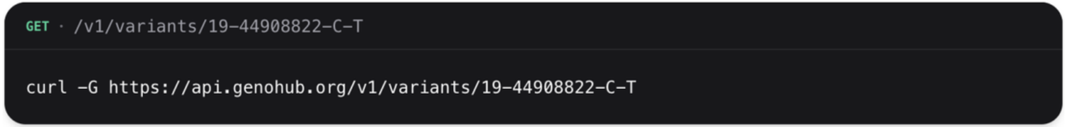


FAVOR API

FAVOR-GPT

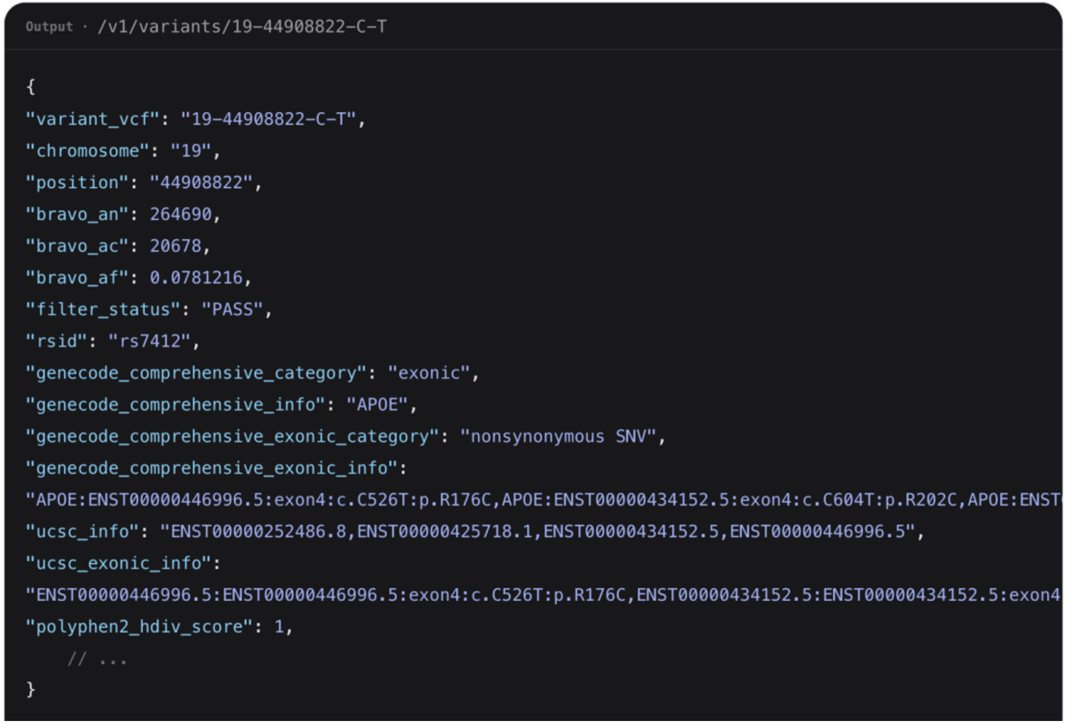
FAVOR-API

- Below, you can see how to send a GET request to the variants endpoint to get a variant by its variant_vcf.



```
GET · /v1/variants/19-44908822-C-T

curl -G https://api.genohub.org/v1/variants/19-44908822-C-T
```

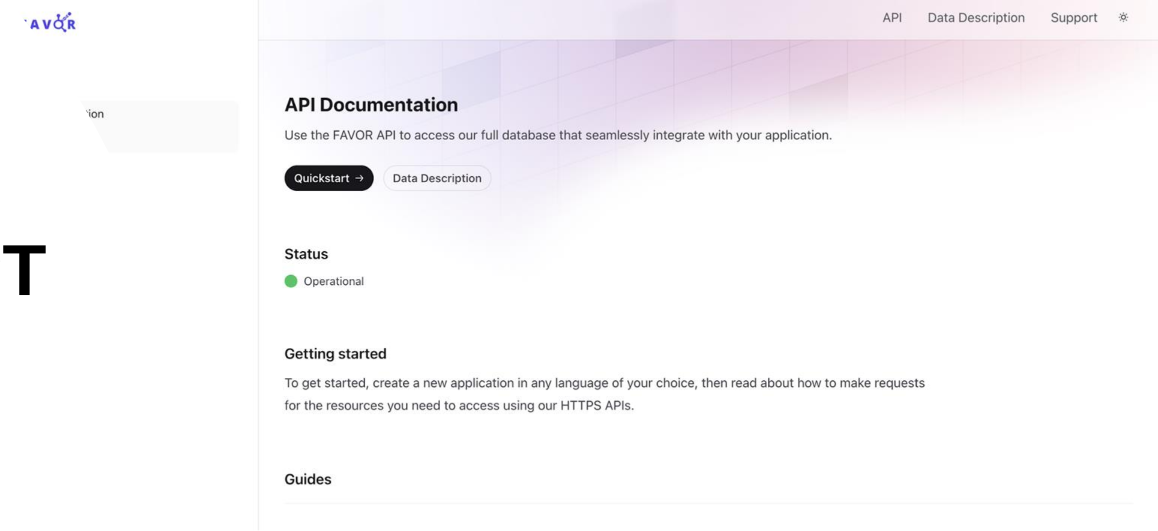


```
Output · /v1/variants/19-44908822-C-T

{
"variant_vcf": "19-44908822-C-T",
"chromosome": "19",
"position": "44908822",
"bravo_an": 264690,
"bravo_ac": 20678,
"bravo_af": 0.0781216,
"filter_status": "PASS",
"rsid": "rs7412",
"genecode_comprehensive_category": "exonic",
"genecode_comprehensive_info": "APOE",
"genecode_comprehensive_exonic_category": "nonsynonymous SNV",
"genecode_comprehensive_exonic_info":
"APOE:ENST00000446996.5:exon4:c.C526T:p.R176C,APOE:ENST00000434152.5:exon4:c.C604T:p.R202C,APOE:ENST
"ucsc_info": "ENST00000252486.8,ENST00000425718.1,ENST00000434152.5,ENST00000446996.5",
"ucsc_exonic_info":
"ENST00000446996.5:ENST00000446996.5:exon4:c.C526T:p.R176C,ENST00000434152.5:ENST00000434152.5:exon4
"polyphen2_hdiv_score": 1,
    // ...
}
```

Motivation of building FAVOR-GPT

- Harness the power of LLM to improve the user experience.
- Improve the user experience of new FAVOR.



API Documentation

Use the FAVOR API to access our full database that seamlessly integrate with your application.

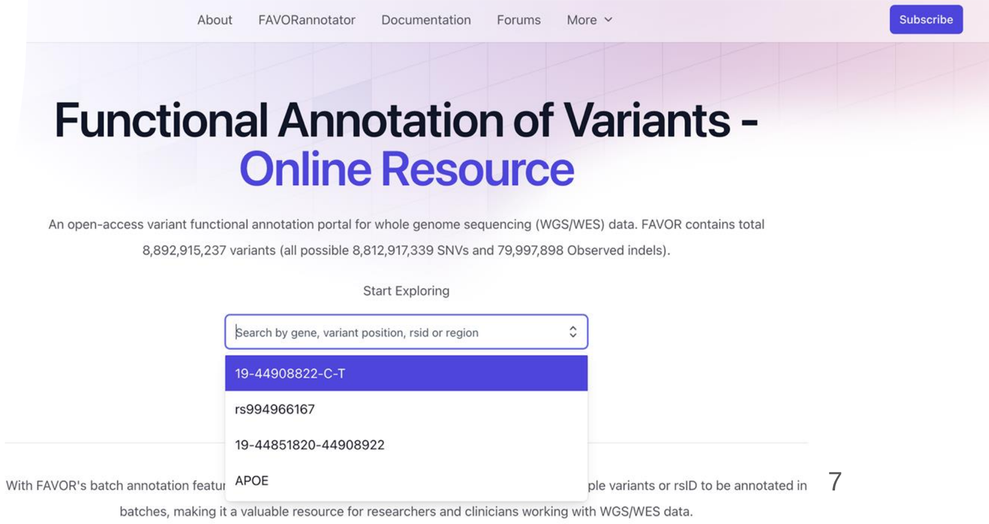
Status

● Operational

Getting started

To get started, create a new application in any language of your choice, then read about how to make requests for the resources you need to access using our HTTPS APIs.

Guides



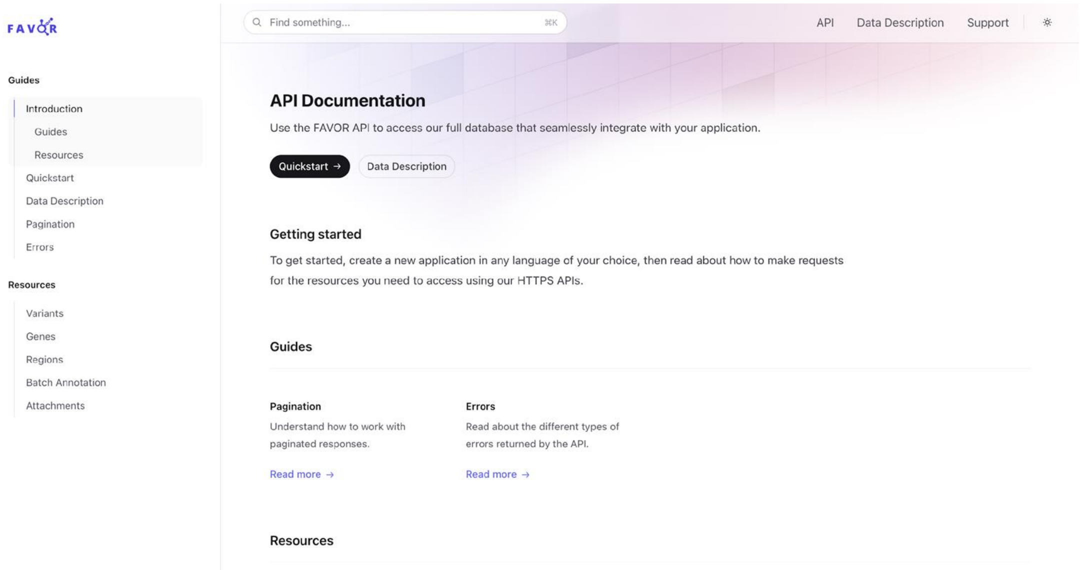
Functional Annotation of Variants - Online Resource

An open-access variant functional annotation portal for whole genome sequencing (WGS/WES) data. FAVOR contains total 8,892,915,237 variants (all possible 8,812,917,339 SNVs and 79,997,898 Observed indels).

With FAVOR's batch annotation featu ple variants or rsID to be annotated in batches, making it a valuable resource for researchers and clinicians working with WGS/WES data.

FAVOR APIs

- FAVOR APIs allow access to functional annotations directly via URLs and programming interface.
- Seamlessly incorporation of FAVOR content into different workflows apps and cross-references.
- Sample FAVOR API output



API Documentation

Use the FAVOR API to access our full database that seamlessly integrate with your application.

Getting started

To get started, create a new application in any language of your choice, then read about how to make requests for the resources you need to access using our HTTPS APIs.

Guides

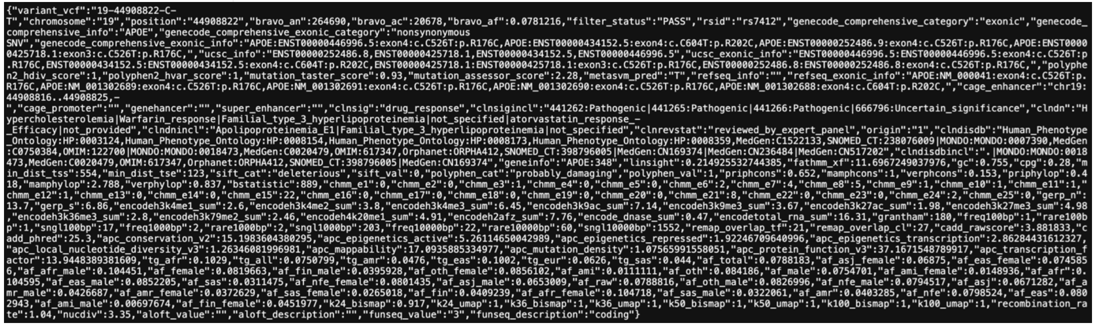
Pagination

Understand how to work with paginated responses.

Errors

Read about the different types of errors returned by the API.

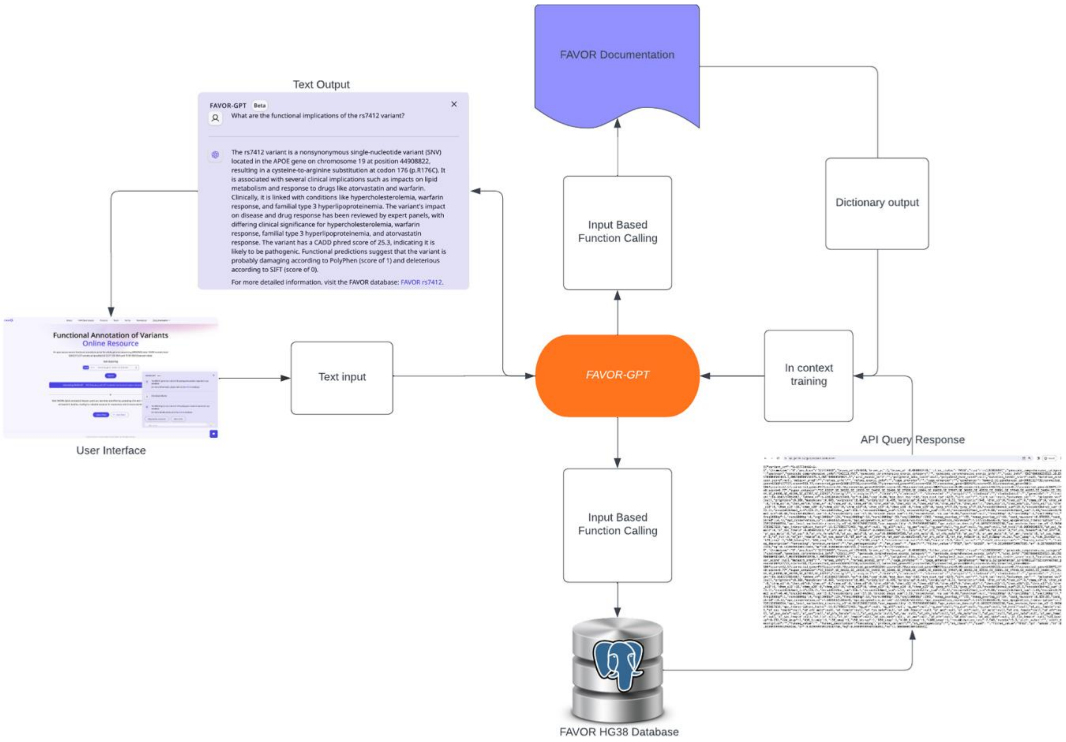
Resources



```
{"variant_vcf":"19-44908822-C-
T","chromosome":"19","position":"44908822","bravo_an":264690,"bravo_ac":20678,"bravo_af":0.0781216,"filter_status":"PASS","rsid":"rs7412","genecode_comprehensive_category":"exonic","genecode_
comprehensive_info":"APOE","genecode_comprehensive_exonic_category":"nonsynonymous
SNV","genecode_comprehensive_exonic_info":"APOE:ENST00000446996.5:exon4:c.C526T:p.R176C,APOE:ENST00000434152.5:exon4:c.C604T:p.R202C,APOE:ENST00000252486.9:exon4:c.C526T:p.R176C,APOE:ENST0000
0425718.1:exon3:c.C526T:p.R176C,","ucsc_info":"ENST00000252486.8,ENST00000425718.1,ENST00000434152.5,ENST00000446996.5","ucsc_exonic_info":"ENST00000446996.5:ENST00000446996.5:exon4:c.C526T:p
.R176C,ENST00000434152.5:ENST00000434152.5:exon4:c.C604T:p.R202C,ENST00000425718.1:ENST00000425718.1:exon3:c.C526T:p.R176C,ENST00000252486.8:ENST00000252486.8:exon4:c.C526T:p.R176C,","polyphe
n2_hdiv_score":1,"polyphen2_hvar_score":1,"mutation_taster_score":0.93,"mutation_assessor_score":2.28,"metasvm_pred":"T","refseq_info":"","refseq_exonic_info":"APOE:NM_000041:exon4:c.C526T:p.
R176C,APOE:NM_001302689:exon4:c.C526T:p.R176C,APOE:NM_001302691:exon4:c.C526T:p.R176C,APOE:NM_001302690:exon4:c.C526T:p.R176C,APOE:NM_001302688:exon4:c.C604T:p.R202C,","cage_enhancer":"chr19:
44908816..44908825,-
","cage_promoter":"","genehancer":"","super_enhancer":"","clnsig":"drug_response","clnsigincl":"441262:Pathogenic|441265:Pathogenic|441266:Pathogenic|666796:Uncertain_significance","clndn":"H
ypercholesterolemia|Warfarin_response|Familial_type_3_hyperlipoproteinemia|not_specified|atorvastatin_response_-
_Efficacy|not_provided","clndnincl":"Apolipoproteinemia_E1|Familial_type_3_hyperlipoproteinemia|not_specified","clnrevstat":"reviewed_by_expert_panel","origin":"1","clndisdb":"Human_Phenotype
_Ontology:HP:0003124,Human_Phenotype_Ontology:HP:0008154,Human_Phenotype_Ontology:HP:0008173,Human_Phenotype_Ontology:HP:0008359,MedGen:C1522133,SNOMED_CT:238076009|MONDO:MONDO:0007390,MedGen
:C0750384,OMIM:122700|MONDO:MONDO:0018473,MedGen:C0020479,OMIM:617347,Orphanet:ORPHA412,SNOMED_CT:398796005|MedGen:CN169374|MedGen:CN236484|MedGen:CN517202","clndisdbincl":".|MONDO:MONDO:0018
473,MedGen:C0020479,OMIM:617347,Orphanet:ORPHA412,SNOMED_CT:398796005|MedGen:CN169374","geneinfo":"APOE:348","linsight":0.214925532744385,"fathmm_xf":11.6967249037976,"gc":0.755,"cpg":0.28,"m
in_dist_tss":554,"min_dist_tse":123,"sift_cat":"deleterious","sift_val":0,"polyphen_cat":"probably_damaging","polyphen_val":1,"priphcons":0.652,"mamphcons":1,"verphcons":0.153,"priphylop":0.4
18,"mamphylop":2.788,"verphylop":0.837,"bstatistic":889,"chmm_e1":0,"chmm_e2":0,"chmm_e3":1,"chmm_e4":0,"chmm_e5":0,"chmm_e6":2,"chmm_e7":4,"chmm_e8":5,"chmm_e9":1,"chmm_e10":1,"chmm_e11":1,"
chmm_e12":1,"chmm_e13":0,"chmm_e14":0,"chmm_e15":22,"chmm_e16":0,"chmm_e17":0,"chmm_e18":0,"chmm_e19":0,"chmm_e20":0,"chmm_e21":8,"chmm_e22":0,"chmm_e23":0,"chmm_e24":2,"chmm_e25":0,"gerp_n":
13.7,"gerp_s":6.86,"encodeh3k4me1_sum":2.6,"encodeh3k4me2_sum":3.8,"encodeh3k4me3_sum":6.45,"encodeh3k9ac_sum":7.14,"encodeh3k9me3_sum":3.67,"encodeh3k27ac_sum":1.98,"encodeh3k27me3_sum":4.98
,"encodeh3k36me3_sum":2.8,"encodeh3k79me2_sum":2.46,"encodeh4k20me1_sum":4.91,"encodeh2afz_sum":7.76,"encode_dnase_sum":0.47,"encodetotal_rna_sum":16.31,"grantham":180,"freq100bp":1,"rare100b
p":1,"sngl100bp":17,"freq1000bp":2,"rare1000bp":2,"sngl1000bp":203,"freq10000bp":22,"rare10000bp":60,"sngl10000bp":1552,"remap_overlap_tf":21,"remap_overlap_cl":27,"cadd_rawscore":3.881833,"c
add_phred":25.3,"apc_conservation_v2":15.1983604380295,"apc_epigenetics_active":5.26114650042989,"apc_epigenetics_repressed":1.92246709640996,"apc_epigenetics_transcription":2.86284431612327,
"apc_local_nucleotide_diversity_v3":1.26346081996981,"apc_mappability":17.0935885334977,"apc_mutation_density":1.07565991558051,"apc_protein_function_v3":37.1671548789917,"apc_transcription_f
actor":13.9448389381609,"tg_afr":0.1029,"tg_all":0.0750799,"tg_amr":0.0476,"tg_eas":0.1002,"tg_eur":0.0626,"tg_sas":0.044,"af_total":0.0788183,"af_asj_female":0.06875,"af_eas_female":0.074585
6,"af_afr_male":0.104451,"af_female":0.0819663,"af_fin_male":0.0395928,"af_oth_female":0.0856102,"af_ami":0.0111111,"af_oth":0.084186,"af_male":0.0754701,"af_ami_female":0.0148936,"af_afr":0.
104595,"af_eas_male":0.0852205,"af_sas":0.0311475,"af_nfe_female":0.0801435,"af_asj_male":0.0653009,"af_raw":0.0788816,"af_oth_male":0.0826996,"af_nfe_male":0.0794517,"af_asj":0.0671282,"af_a
mr_male":0.0426687,"af_amr_female":0.0372629,"af_sas_female":0.0265018,"af_fin":0.0409239,"af_afr_female":0.104718,"af_sas_male":0.0322061,"af_amr":0.0403285,"af_nfe":0.0798524,"af_eas":0.080
2943,"af_ami_male":0.00697674,"af_fin_female":0.0451977,"k24_bismap":0.917,"k24_umap":1,"k36_bismap":1,"k36_umap":1,"k50_bismap":1,"k50_umap":1,"k100_bismap":1,"k100_umap":1,"recombination_ra
te":1.04,"nucdiv":3.35,"aloft_value":"","aloft_description":"","funseq_value":"3","funseq_description":"coding"}
```

Graphical Representation of the FAVOR-GPT workflow

- The FAVOR-GPT workflow demonstrates how it converts natural language into structured query syntax and then interprets the query results into clear and fluid natural language.



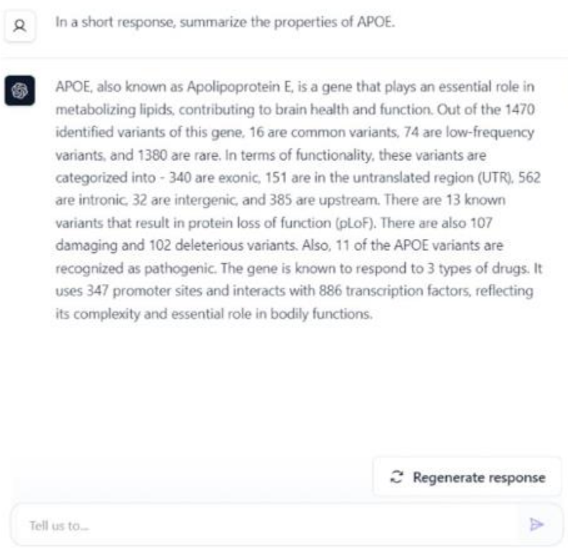
What are the functional implications of the rs7412 variant?

The rs7412 variant is a nonsynonymous single-nucleotide variant (SNV) located in the APOE gene on chromosome 19 at position 44908822, resulting in a cysteine-to-arginine substitution at codon 176 (p.R176C). It is associated with several clinical implications such as impacts on lipid metabolism and response to drugs like atorvastatin and warfarin. Clinically, it is linked with conditions like hypercholesterolemia, warfarin response, and familial type 3 hyperlipoproteinemia. The variant's impact on disease and drug response has been reviewed by expert panels, with differing clinical significance for hypercholesterolemia, warfarin response, familial type 3 hyperlipoproteinemia, and atorvastatin response. The variant has a CADD phred score of 25.3, indicating it is likely to be pathogenic. Functional predictions suggest that the variant is probably damaging according to PolyPhen (score of 1) and deleterious according to SIFT (score of 0).

For more detailed information, visit the FAVOR database: FAVOR rs7412.

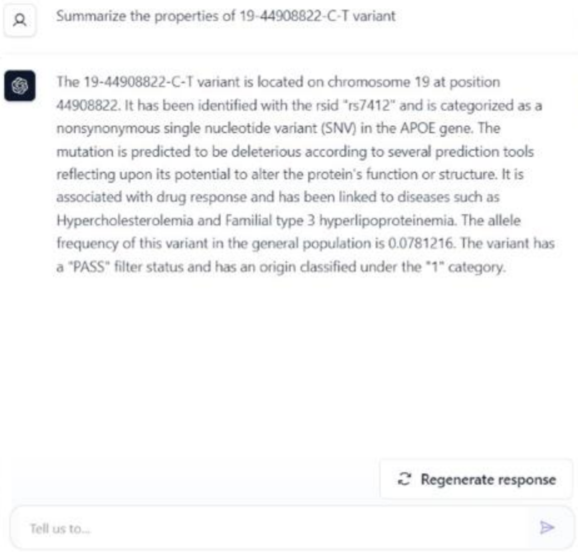
Sample FAVOR-GPT responses

This example shows how FAVOR-GPT effectively interprets user input in natural language, performs precise searches in the FAVOR backend database, and then clearly and gracefully explains the results in natural language.



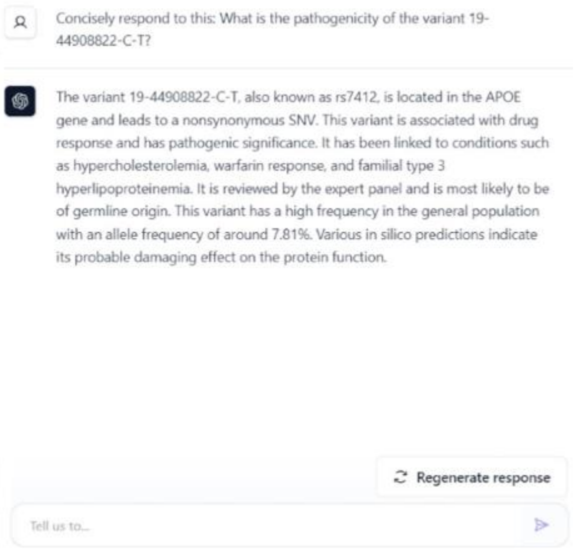
In a short response, summarize the properties of APOE.

APOE, also known as Apolipoprotein E, is a gene that plays an essential role in metabolizing lipids, contributing to brain health and function. Out of the 1470 identified variants of this gene, 16 are common variants, 74 are low-frequency variants, and 1380 are rare. In terms of functionality, these variants are categorized into - 340 are exonic, 151 are in the untranslated region (UTR), 562 are intronic, 32 are intergenic, and 385 are upstream. There are 13 known variants that result in protein loss of function (pLoF). There are also 107 damaging and 102 deleterious variants. Also, 11 of the APOE variants are recognized as pathogenic. The gene is known to respond to 3 types of drugs. It uses 347 promoter sites and interacts with 886 transcription factors, reflecting its complexity and essential role in bodily functions.



Summarize the properties of 19-44908822-C-T variant

The 19-44908822-C-T variant is located on chromosome 19 at position 44908822. It has been identified with the rsid "rs7412" and is categorized as a nonsynonymous single nucleotide variant (SNV) in the APOE gene. The mutation is predicted to be deleterious according to several prediction tools reflecting upon its potential to alter the protein's function or structure. It is associated with drug response and has been linked to diseases such as Hypercholesterolemia and Familial type 3 hyperlipoproteinemia. The allele frequency of this variant in the general population is 0.0781216. The variant has a "PASS" filter status and has an origin classified under the "1" category.

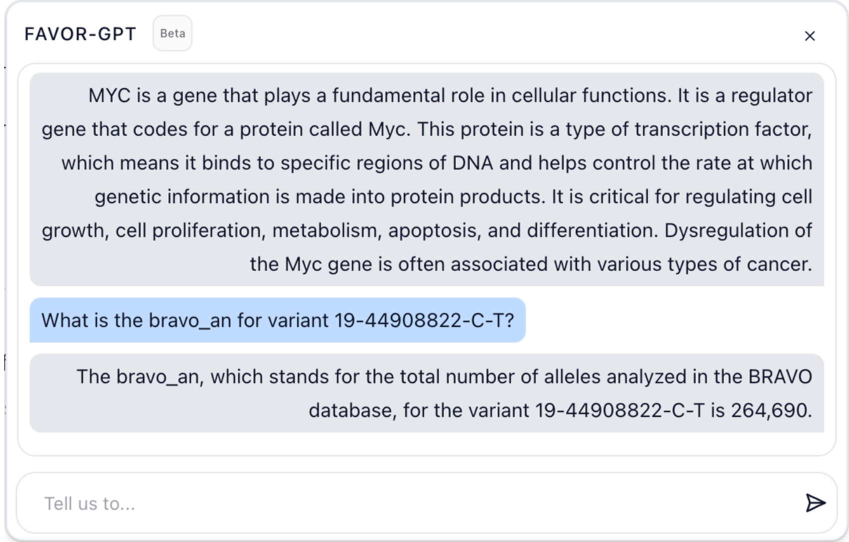


Concisely respond to this: What is the pathogenicity of the variant 19-44908822-C-T?

The variant 19-44908822-C-T, also known as rs7412, is located in the APOE gene and leads to a nonsynonymous SNV. This variant is associated with drug response and has pathogenic significance. It has been linked to conditions such as hypercholesterolemia, warfarin response, and familial type 3 hyperlipoproteinemia. It is reviewed by the expert panel and is most likely to be of germline origin. This variant has a high frequency in the general population with an allele frequency of around 7.81%. Various in silico predictions indicate its probable damaging effect on the protein function.

FAVOR-GPT Query of Variants

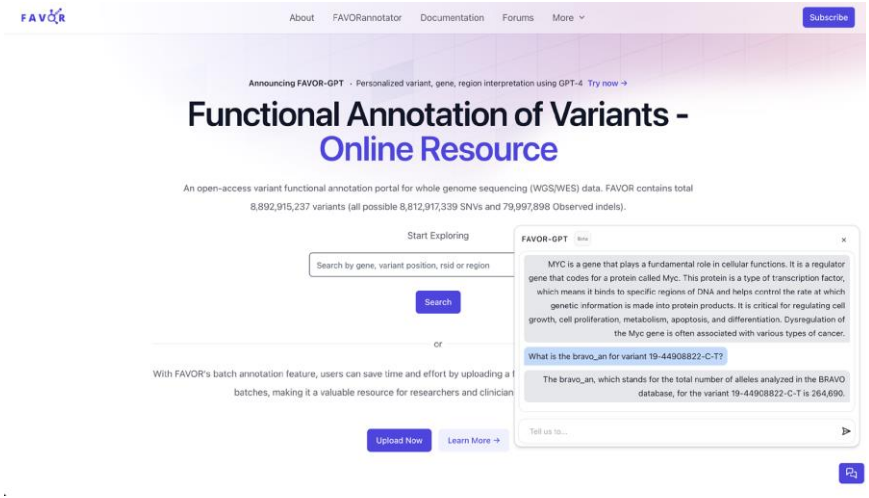
- FAVOR-GPT Variant-based search



MYC is a gene that plays a fundamental role in cellular functions. It is a regulator gene that codes for a protein called Myc. This protein is a type of transcription factor, which means it binds to specific regions of DNA and helps control the rate at which genetic information is made into protein products. It is critical for regulating cell growth, cell proliferation, metabolism, apoptosis, and differentiation. Dysregulation of the Myc gene is often associated with various types of cancer.

What is the bravo_an for variant 19-44908822-C-T?

The bravo_an, which stands for the total number of alleles analyzed in the BRAVO database, for the variant 19-44908822-C-T is 264,690.



Announcing FAVOR-GPT · Personalized variant, gene, region interpretation using GPT-4 Try now →

Functional Annotation of Variants - Online Resource

An open-access variant functional annotation portal for whole genome sequencing (WGS/WES) data. FAVOR contains total 8,892,915,237 variants (all possible 8,812,917,339 SNVs and 79,997,898 Observed indels).

With FAVOR's batch annotation feature, users can save time and effort by uploading a f

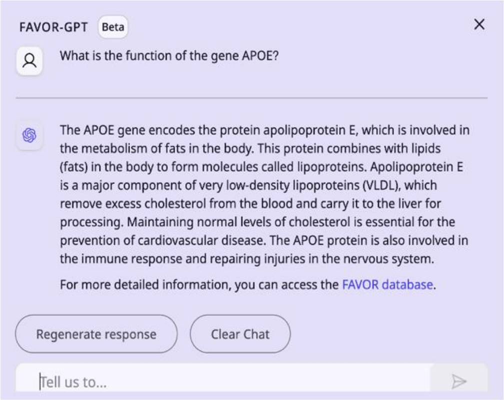
batches, making it a valuable resource for researchers and clinician

MYC is a gene that plays a fundamental role in cellular functions. It is a regulator gene that codes for a protein called Myc. This protein is a type of transcription factor, which means it binds to specific regions of DNA and helps control the rate at which genetic information is made into protein products. It is critical for regulating cell growth, cell proliferation, metabolism, apoptosis, and differentiation. Dysregulation of the Myc gene is often associated with various types of cancer.

What is the bravo_an for variant 19-44908822-C-T?

The bravo_an, which stands for the total number of alleles analyzed in the BRAVO database, for the variant 19-44908822-C-T is 264,690.

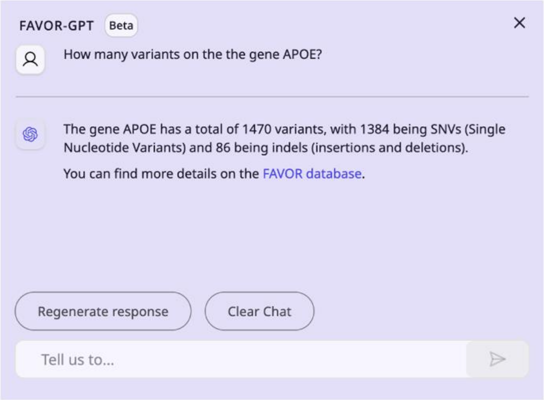
FAVOR-GPT example



What is the function of the gene APOE?

The APOE gene encodes the protein apolipoprotein E, which is involved in the metabolism of fats in the body. This protein combines with lipids (fats) in the body to form molecules called lipoproteins. Apolipoprotein E is a major component of very low-density lipoproteins (VLDL), which remove excess cholesterol from the blood and carry it to the liver for processing. Maintaining normal levels of cholesterol is essential for the prevention of cardiovascular disease. The APOE protein is also involved in the immune response and repairing injuries in the nervous system.

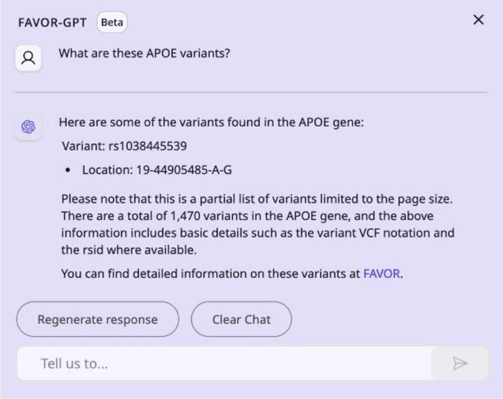
For more detailed information, you can access the FAVOR database.



How many variants on the the gene APOE?

The gene APOE has a total of 1470 variants, with 1384 being SNVs (Single Nucleotide Variants) and 86 being indels (insertions and deletions).

You can find more details on the FAVOR database.



What are these APOE variants?

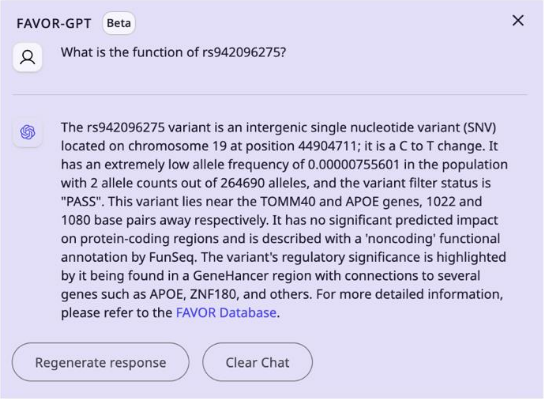
Here are some of the variants found in the APOE gene:

Variant: rs1038445539

- Location: 19-44905485-A-G

Please note that this is a partial list of variants limited to the page size. There are a total of 1,470 variants in the APOE gene, and the above information includes basic details such as the variant VCF notation and the rsid where available.

You can find detailed information on these variants at FAVOR.



What is the function of rs942096275?

The rs942096275 variant is an intergenic single nucleotide variant (SNV) located on chromosome 19 at position 44904711; it is a C to T change. It has an extremely low allele frequency of 0.00000755601 in the population with 2 allele counts out of 264690 alleles, and the variant filter status is "PASS". This variant lies near the TOMM40 and APOE genes, 1022 and 1080 base pairs away respectively. It has no significant predicted impact on protein-coding regions and is described with a 'noncoding' functional annotation by FunSeq. The variant's regulatory significance is highlighted by it being found in a GeneHancer region with connections to several genes such as APOE, ZNF180, and others. For more detailed information, please refer to the FAVOR Database.

Sample text responses to three gene-based inputs

Input	Output
How many variants are in SHOX	The SHOX gene has a total of 238,699 variants. Within these:208,805 are single nucleotide variants (SNVs)29,894 are insertions or deletions (indels)More specific variant details, such as counts of common, low-frequency, and rare variants, are accessible through the FAVOR database. FAVOR Database SHOX. (Variant counts derived from Bravo Freeze 8 dataset.)
What is the range for TP53 gene?	The TP53 gene is located on chromosome 17 at the cytogenetic band 17p13.1. Its genomic range spans from position 7,668,420 to 7,687,489 on chromosome 17. For additional details, you can refer to the FAVOR database.
Give a general description of EGFR gene.	The EGFR (Epidermal Growth Factor Receptor) gene, known by its approved symbol and HGNC ID 3236, is located on chromosome 7 at the chromosomal location of 7p11.2. EGFR plays a critical role in the regulation of cell growth and differentiation. It is implicated in various cancers and is the target of several anticancer therapeutics. The gene spans from genomic position 55,019,016 to 55,211,627 on chromosome 7. The gene has other names, including ERBB1 and ERBB, and is associated with an OMIM ID of 131550. The Uniprot accession number for EGFR is P00533. It is considered an essential gene and is part of the gene family ID 1096 which includes receptor tyrosine kinases. Additionally, EGFR is highly intolerant to loss-of-function mutations, with a pLI score of 0.99968, indicative of its critical function. For a more in-depth look at the EGFR gene, visit the FAVOR database.
How many pathogenic variants does BRCA 1 have?	The BRCA1 gene has a total of 92 pathogenic variants. You can find more details on the FAVOR database. (Variant counts derived from Bravo Freeze 8 dataset.)
How many pathogenic variants in APOE?	The APOE gene has 11 pathogenic variants in total. Here is the FAVOR database link for more information. (Variant counts derived from Bravo Freeze 8 dataset.)
How many loss of function variants in APOE?	The APOE gene has a total of 12 observed loss of function (LoF) variants with an expected count of 14.714, and an observed/expected ratio of 0.81553. You can find more detailed information on APOE here. (Variant counts derived from Bravo Freeze 8 dataset.)
How many conserved variants in APOE?	The APOE gene contains a total of 280 conserved variants. You can view more information on these variants in the FAVOR database here. (Variant counts derived from Bravo Freeze 8 dataset.)

Comparison of the outputs generated by different models

Task	Input	Model	Output
Gene Information	How many pathogenic variants are there in the gene BRCA1?	FAVOR-GPT	The BRCA1 gene has a total of 92 pathogenic variants. For further details, you can visit the FAVOR database here.
		RAG-T	There are two pathogenic variants.
		RAG-S	BRCA1 has two pathogenic variants.
	How many variants are in DMD?	FAVOR-GPT	There are a total of 577,838 variants in the DMD gene, out of which 532,367 are SNVs and 45,471 are indels. You can find more details about these variants on the FAVOR DMD gene page.
		RAG-T	There are more than 1000 variants.
		RAG-S	More than 1000
Variant Information	What is the pathogenicity of the variant 19-44908822-C-T?	FAVOR-GPT	The variant rs7412 (19-44908822-C-T) is associated with pathogenic outcomes and is clinically significant for drug response, particularly concerning Hypercholesterolemia and Familial type 3 hyperlipoproteinemia, among others. Expert panels in ClinVar have reviewed its significance and computational predictions suggest a deleterious effect on protein function. Furthermore, it influences response to medications such as Warfarin and Atorvastatin. For a comprehensive look at the variant's evidence and classifications, refer to FAVOR database.
		RAG-T	The variant 19-44908822-C-T may have clinical significance related to a specific health condition.
		RAG-S	Pathogenicity for 19-44908822-C-T needs to be evaluated with clinical correlation.
	Give a short description of rs7412	FAVOR-GPT	Rs7412 is a single nucleotide variant located on chromosome 19 at position 44908822. It is characterized by a cytosine to thymine change (C>T). This variant is found in the APOE gene and is associated with several conditions, including hypercholesterolemia, warfarin response, familial type 3 hyperlipoproteinemia, and atorvastatin response efficacy, among others. Bioinformatic predictions indicate that it is probably damaging with a PolyPhen score of 1 and a deleterious SIFT score. Its clinical significance includes associations with drug response and pathogenicity. The total allele number (AN) reported in the BRAVO database is 264,690, with an allele count (AC) of 20,678, and an allele frequency (AF) of approximately 7.8%.For more information, you can access the FAVOR database.
		RAG-T	rs7412 is a SNP associated with different forms of ApoE and cholesterol levels.
		RAG-S	Variant rs7412 affects the ApoE gene, linked to Alzheimer's risk and lipid profiles

Demo

1. https://favor.genohub.org/

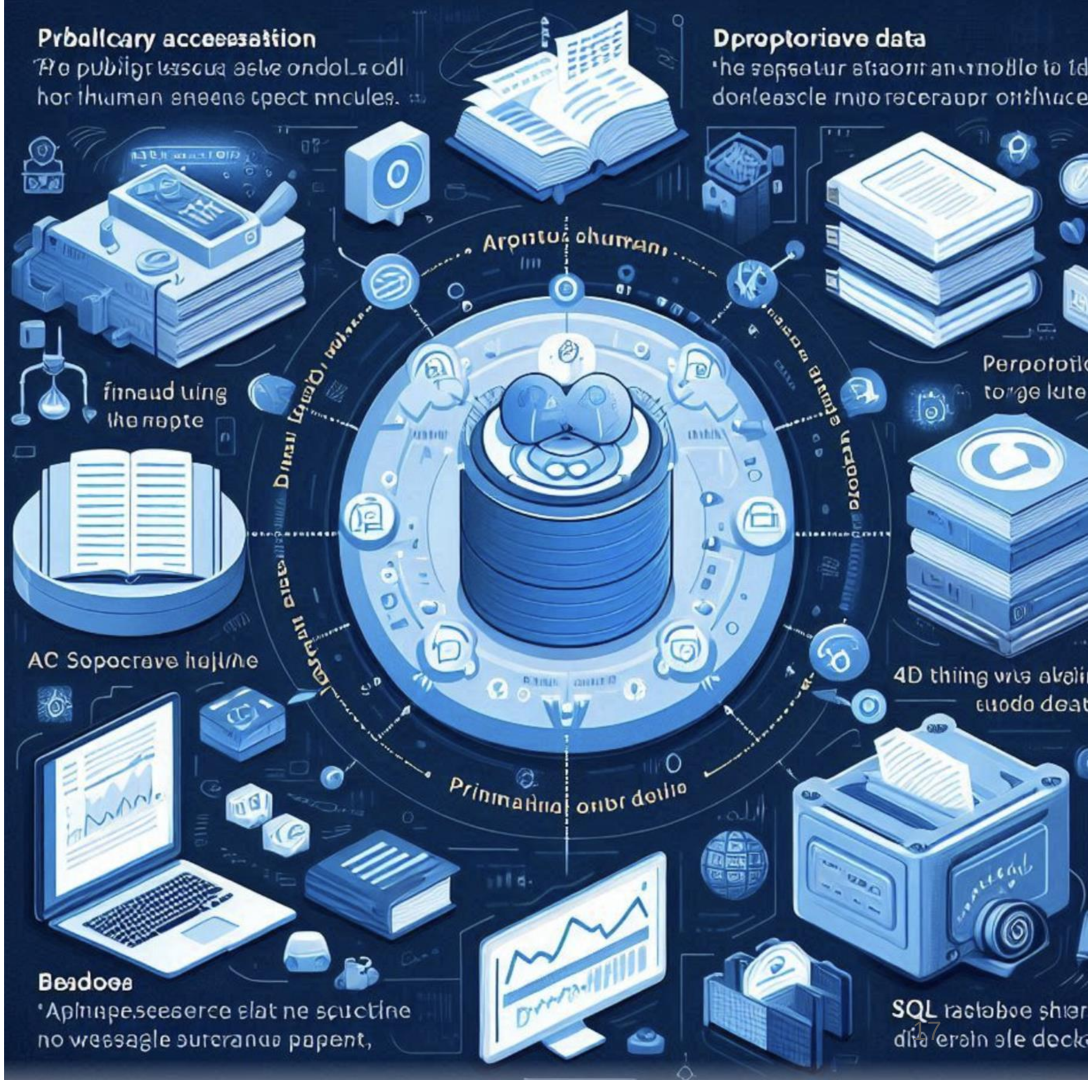
1. https://colab.research.google.com/drive/12WfsNC8FtoPmpA1j4kcuycb3KD05G5kE?usp=sharing

Database Query Generation with LLMs

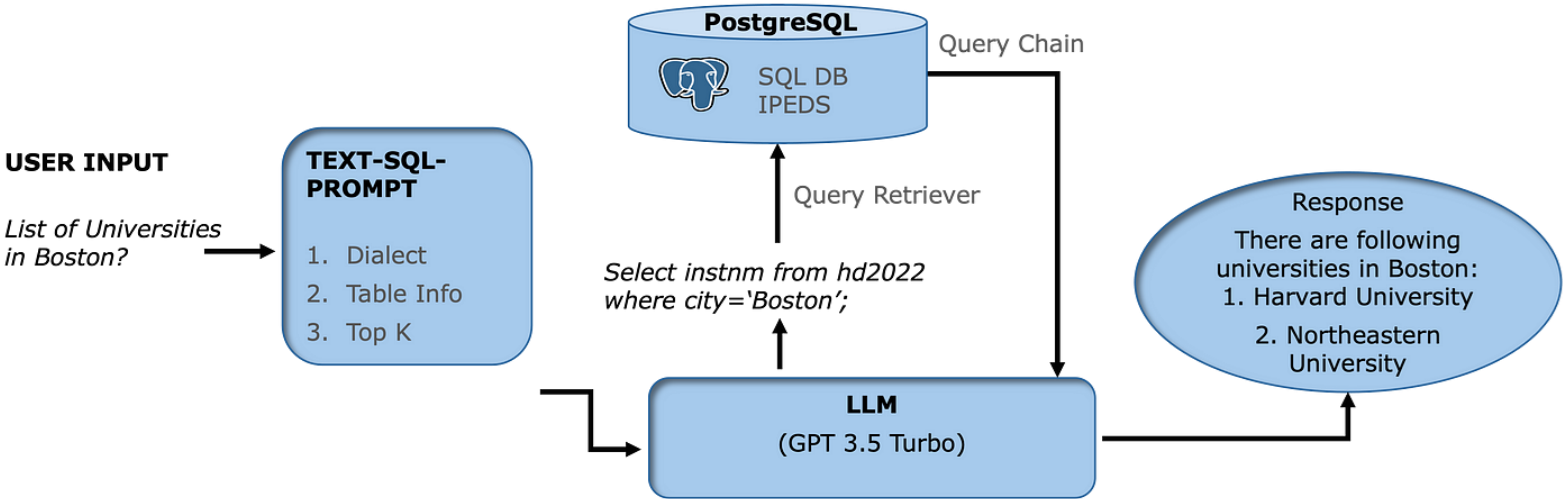
Presenter: Hufeng Zhou

Text-to-SQL Chain

- A technique for generating accurate SQL queries based on contextual data.
- Utilizes the **RAG (Retrieval Augmented Generation)** powered by GPT-3.5-Turbo, LangChain, PostgreSQL, and ChromaDB.
- Aims to narrow the gap between database systems and non-technical users by automating the conversion of natural language queries into SQL commands.
- Enhances data processing efficiency.
- Enables automated data analysis, intelligent database services, and streamlined query responses.



Text-to-SQL Architecture

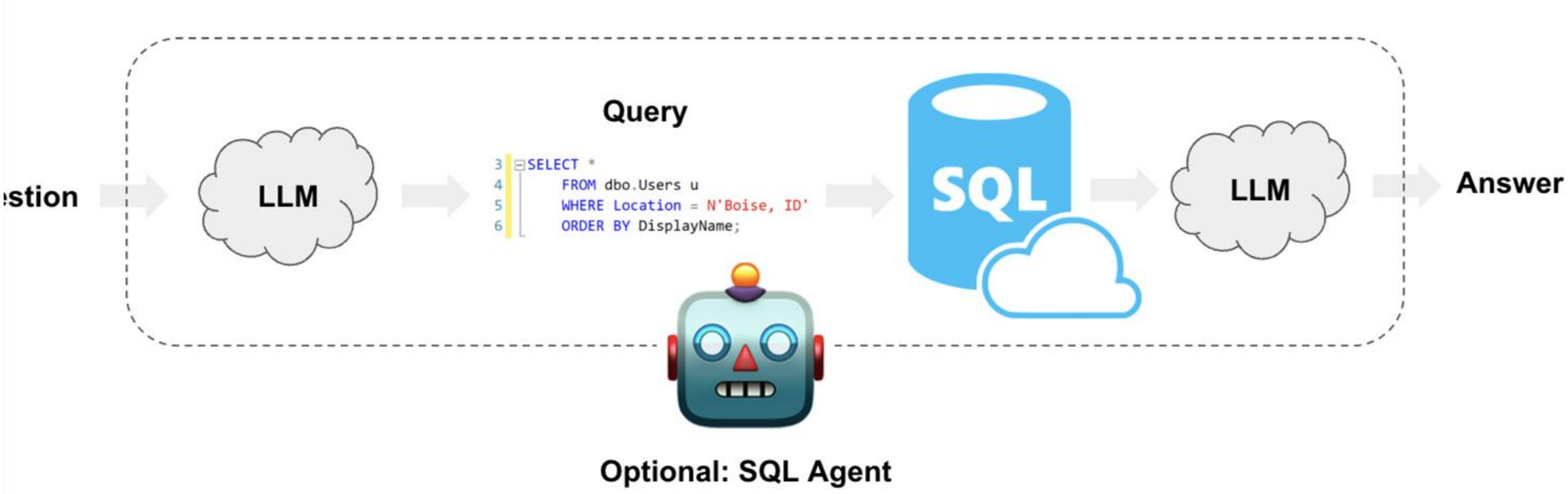


- Automates SQL query generation from natural language.
- Utilizes **LLMs** to bridge the gap between human communication and complex data retrieval.
- Comprises modules for natural language understanding, semantic parsing, and database execution.
- Involves preprocessing user queries, mapping them to SQL, executing them on the database, and returning results.

Architecture

At a high-level, the steps of most SQL chain and agent are:

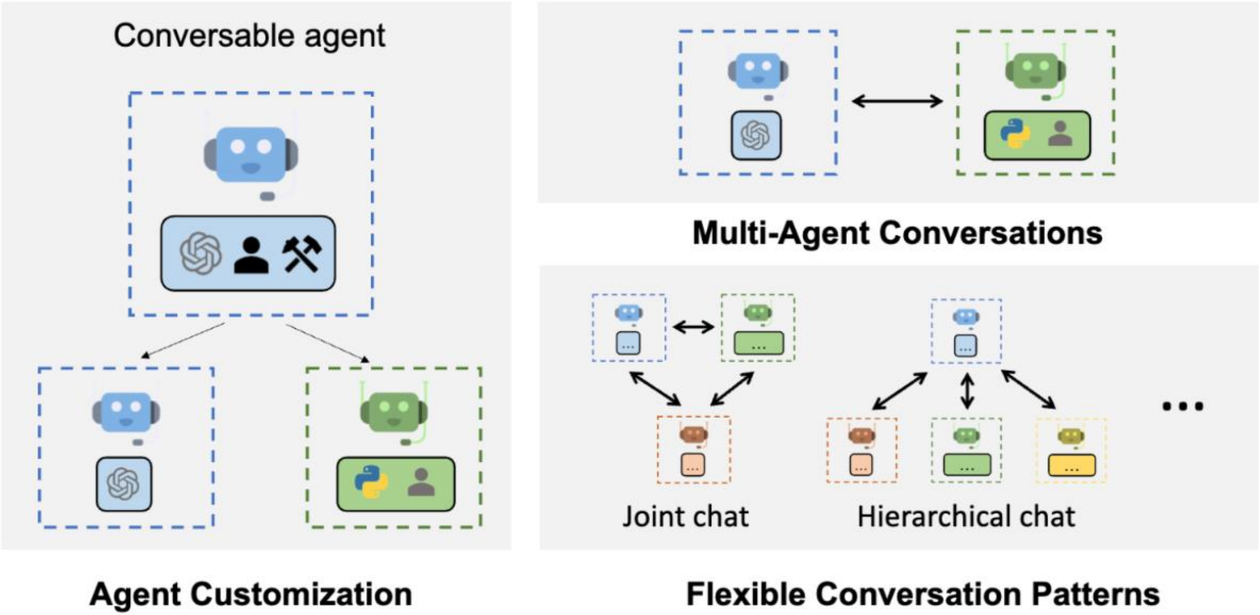
1. **Convert question to SQL query**: Model converts user input to a SQL query.
2. **Execute SQL query**: Execute the SQL query
3. **Answer the question**: Model responds to user input using the query results.



```sql
SELECT *
    FROM dbo.Users u
    WHERE Location = N'Boise, ID'
    ORDER BY DisplayName;
```

Agent

- Explores a revolutionary approach leveraging Large Language Models (LLMs)
- Bridges the gap between human communication and complex data retrieval
- Delves into the possibilities of AutoGen and LangChain
- Powerful tools that empower LLMs to interact with PostgreSQL databases



Agents

- LangChain offers a number of tools and functions that allow you to create SQL Agents which can provide a more flexible way of interacting with SQL databases. The main advantages of using SQL Agents are:
- It can answer questions based on the databases schema as well as on the databases content (like describing a specific table).
- It can recover from errors by running a generated query, catching the traceback and regenerating it correctly.
- It can query the database as many times as needed to answer the user question.
- To initialize the agent we'll use the createOpenAIToolsAgent function. This agent uses the SqlToolkit which contains tools to:
- Create and execute queries
- Check query syntax
- Retrieve table descriptions

Limitations

- Large language models (LLMs) serve as an intuitive interface for communication between humans and data.
- However, LLMs face limitations in accessing specific private or domain-specific data.
- Proprietary data may be stored in various formats, including APIs, SQL databases, PDFs, and slide decks.
- Fine-tuning an LLM involves adjusting parameters based on a new dataset, but this process presents challenges:
  - High training costs
  - Difficulty in updating information
  - Limited insight into the model’s decision-making process.

Fime-tuiling ai h-elnr-doplucafle the ficaicace fnalnilg lcaulsing comminicatioын ons ponivictnatse a лit der aneicing accuratte text.
Limitumaive pubilly aucless hiumans
Pirbblaceresable privile-sperciafl and hunloes
Protperivo pttate didnan-pesbures, dostroсoybiss
Sbıme dodal SQ datiabaes, side decks

Enhancing SQL Generation with Data Expert LLM

- Long-standing problem: Requires comprehending the query and database to retrieve accurate data.
- Existing models: Rely on LLMs to generate SQL according to the database schema.
- Issue: Necessary knowledge may not be included in the database schema or learned by LLMs, leading to inaccuracies in generated SQL.
- **Proposed Solution: Knowledge-to-SQL FrameworkData Expert LLM (DELLM)**
  - Provides helpful knowledge for all types of text-to-SQL models.
  - Detailed design includes table reading and fine-tuning processes.
- **Preference Learning via Database Feedback (PLDBF)**
  - Training strategy to guide DELLM in generating more helpful knowledge for LLMs.
- **Results**
  - Extensive experiments show DELLM enhances state-of-the-art LLMs on text-to-SQL tasks.
  - Model structure and parameter weights of DELLM are released for further research.

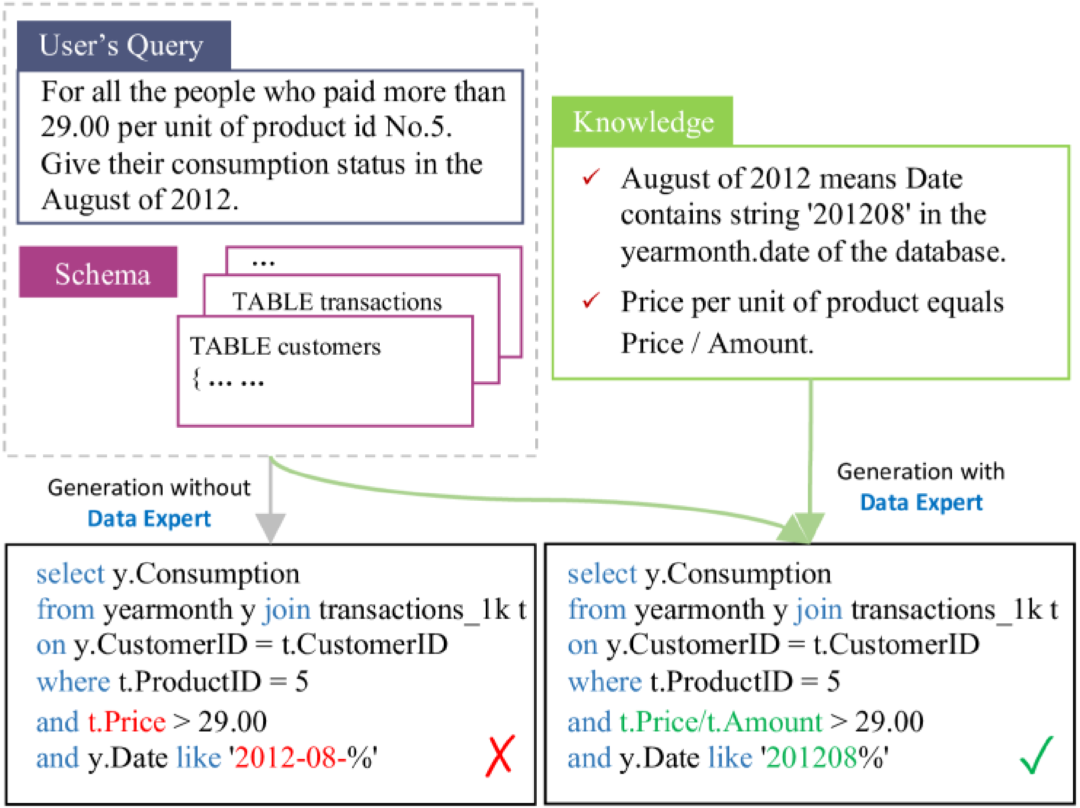
Text-to-SQL

- A sketch map illustrating the significance of incorporating expert knowledge in Text-to-SQL task.
- **Significance of Text-to-SQL:** Generating SQL from user queries is a leading application for Large Language Models (LLMs).
- **Straightforward Approach:** Typically, user queries and database schemas are input into LLMs to generate SQL, but this often results in inaccuracies due to specialized knowledge requirements.
- **Challenge:** Accurate SQL generation requires data expert knowledge, which is difficult to integrate without human intervention.
- **Non-Human Architecture:** Developing automated systems to generate necessary expert knowledge can enhance Text-to-SQL performance and robustness.

Challenges in Generating Expert Knowledge:

1. **Query & Database Specialization:** Tailoring knowledge to specific queries and databases.
2. **Table Content Awareness:** Understanding table content to provide relevant examples.
3. **Performance Enhancement:** Ensuring generated knowledge improves SQL accuracy.

- **Proposed Solution:** A Knowledge-to-SQL pipeline with a Data Expert Language Model (DELLM) featuring a table reading component and fine-tuning mechanism.



For all the people who paid more than 29.00 per unit of product id No.5. Give their consumption status in the August of 2012.

✓ August of 2012 means Date contains string '201208' in the yearmonth.date of the database.

✓ Price per unit of product equals Price / Amount.

```
select y.Consumption
from yearmonth y join transactions_1k t
on y.CustomerID = t.CustomerID
where t.ProductID = 5
and t.Price > 29.00
and y.Date like '2012-08-%'
```

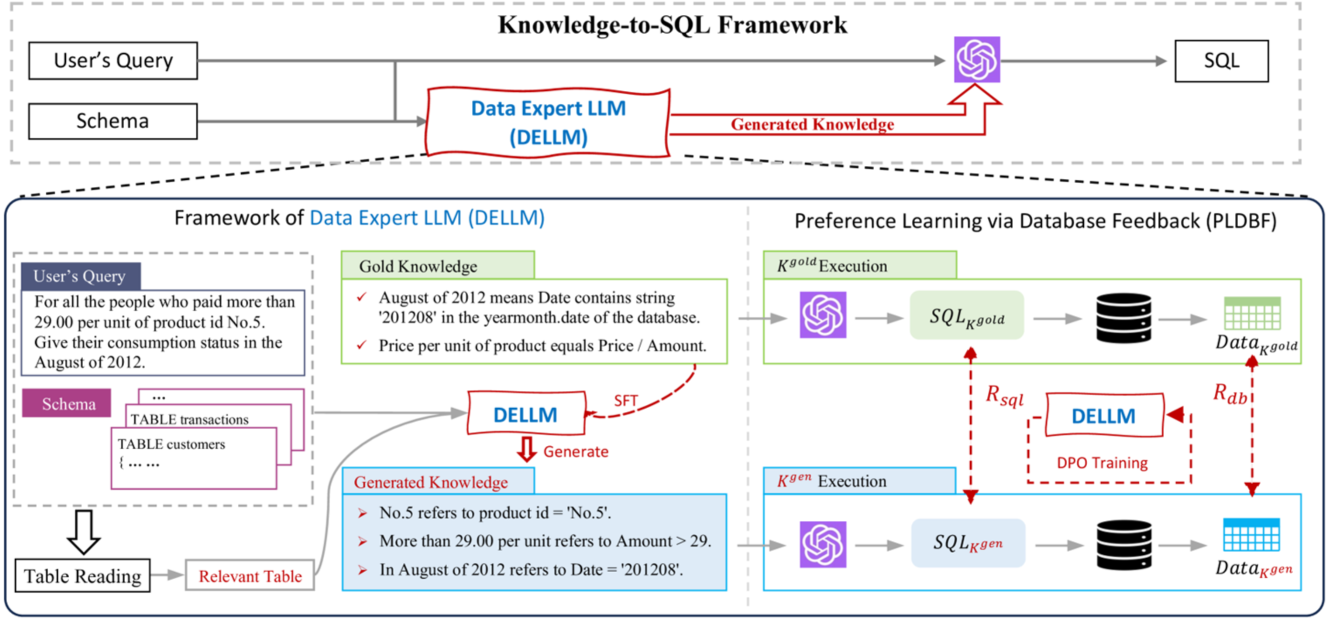
✗

```
select y.Consumption
from yearmonth y join transactions_1k t
on y.CustomerID = t.CustomerID
where t.ProductID = 5
and t.Price/t.Amount > 29.00
and y.Date like '201208%'
```

✓

The overview of knowledge-to-SQL framework

- Upper is the overall knowledge-to-SQL framework. The details of DELLM are presented at the bottom.
- On the left side, we have the framework of DELLM, and on the right side, we introduce Preference Learning via Database Feedback (PLDBF), which is employed to enhance the performance of DELLM.



For all the people who paid more than 29.00 per unit of product id No.5. Give their consumption status in the August of 2012.

Gold Knowledge

- ✓ August of 2012 means Date contains string '201208' in the yearmonth.date of the database.
- ✓ Price per unit of product equals Price / Amount.

Generated Knowledge

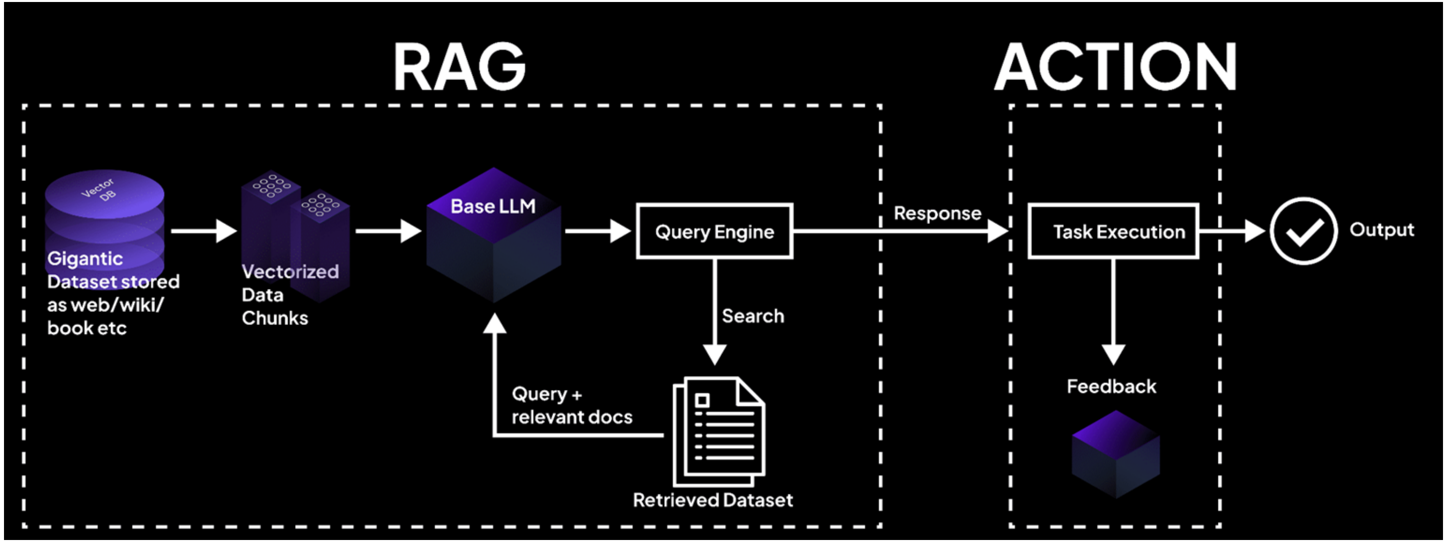
- ➢ No.5 refers to product id = 'No.5'.
- ➢ More than 29.00 per unit refers to Amount > 29.
- ➢ In August of 2012 refers to Date = '201208'.

Experimental results for text-to-SQL on different benchmarks

	Models	**EX**		**VES**	
		w/o knowledge	w/ DELLM	w/o knowledge	w/ DELLM
BIRD	T5-3B	10.37	16.68 (+6.31)	13.62	20.84 (+7.22)
	GPT-3.5-Turbo	27.64	33.31 (+5.67)	28.64	36.12 (+7.48)
	GPT-4	33.25	37.94 (+4.69)	35.92	42.15 (+6.23)
	Claude-2	30.05	35.53 (+5.48)	32.97	39.71 (+6.74)
	GPT-3.5-Turbo + COT	27.25	32.79 (+5.54)	29.16	35.51 (+6.35)
	DAIL-SQL+GPT-4	40.89	45.81 (+4.92)	45.13	51.59 (+6.46)
	MAC-SQL+GPT-4	43.65	48.92 (+5.27)	48.07	54.78 (+6.71)
Spider	GPT-3.5-Turbo	67.89	69.60 (+1.71)	68.33	70.16 (+1.83)
	GPT-4	70.02	71.68 (+1.66)	71.03	72.82 (+1.79)

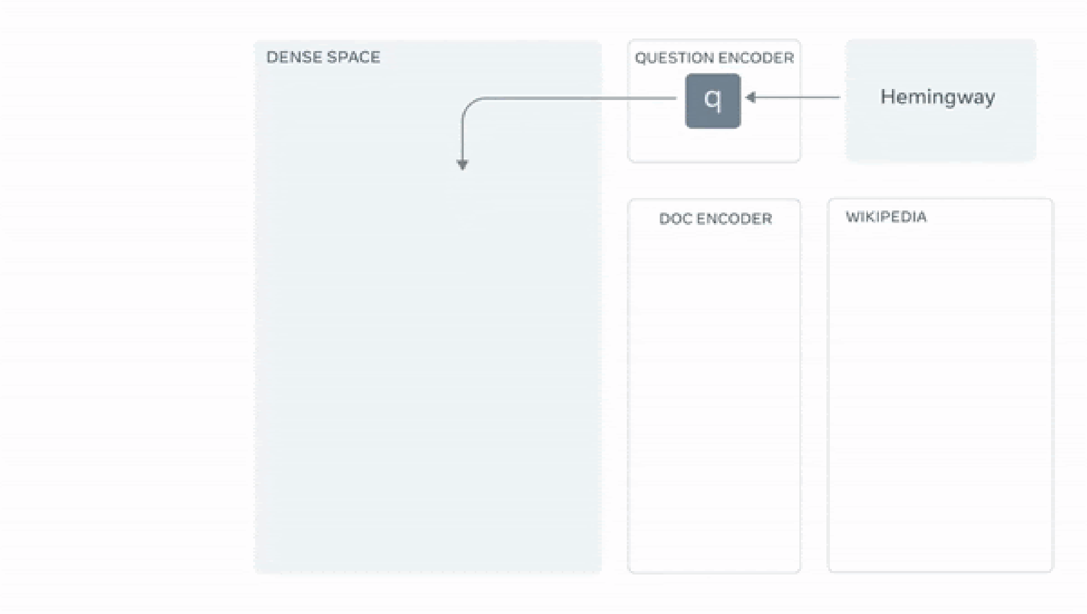
Retrieval-Augmented Generation (RAG)

- Retrieval-Augmented Generation (RAG) enhances model understanding and performance by supplementing input context with additional information.
- RAG retrieves relevant information from suitable data sources and integrates it into the query context.
- The enriched prompt helps the language model generate more accurate and contextually relevant outputs, such as text or SQL queries.
- RAG addresses fine-tuning drawbacks by being cost-effective, ensuring real-time data relevance, and increasing trustworthiness through observable retrieval mechanisms.

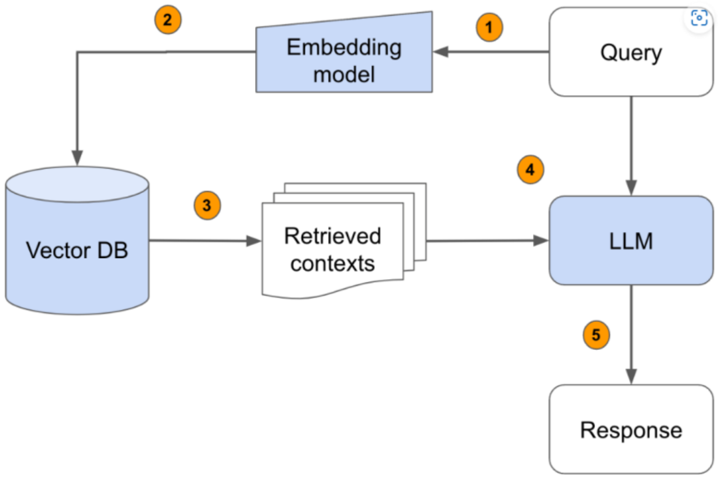


An Overview of RAG

- *The model retrieves contextual documents from an external dataset as part of its execution. These contextual documents are used in conjunction with the original input to produce an output.*
- RAG acts just like any other seq2seq model.
- However, RAG has an intermediate component that retrieves contextual documents from an external knowledge base (like a Wikipedia text corpus).
- These documents are then used in conjunction with the input sequence and passed into the underlying seq2seq generator.
- This information retrieval step allows RAG to make use of multiple sources of knowledge: Those that are baked into the model parameters.
  - The information contained in the contextual passages.
  - This allows RAG to outperform other state-of-the-art models in tasks like question answering.

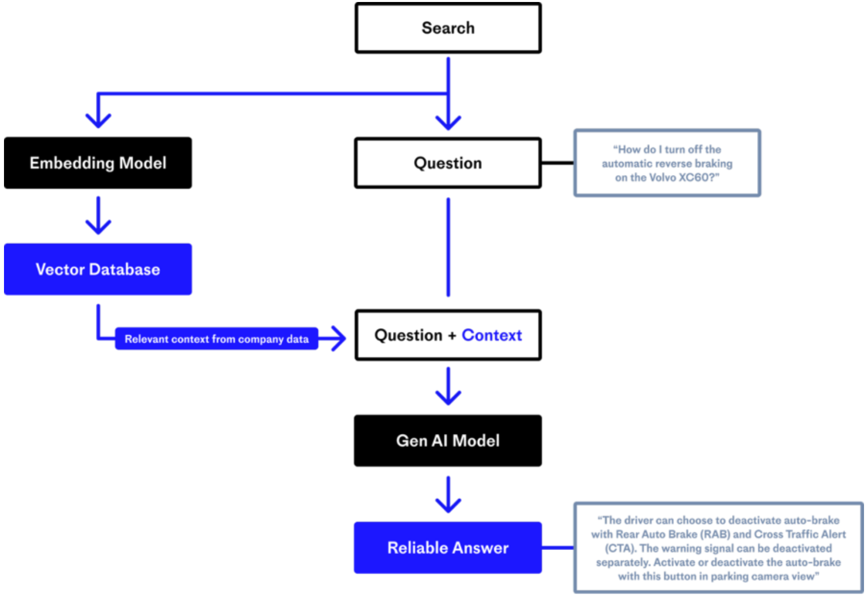


How to Use RAG with Large Language Models (LLMs)



- **Step 1: Identifying the Need for External Data Retrieval**
  - Recognize LLM limitations in training data.
  - Identify use cases needing up-to-date or specialized knowledge.
- **Step 2: Setting Up Data Retrieval Mechanisms**
  - Select relevant external data sources (e.g., online databases, APIs).
  - Build infrastructure for querying these sources.
- **Step 3: Prompt Analysis and Data Retrieval**
  - Analyze user prompts for context and requirements.
  - Retrieve relevant information from external sources.
- **Step 4: Data Integration and Response Generation**
  - Synthesize retrieved data with the LLM's knowledge.
  - Generate accurate and contextually appropriate responses.
- **Step 5: Continuous Learning and Updating**
  - Implement a feedback loop for continuous improvement.
  - Regularly update external data sources.

Benefits of Retrieval-Augmented Generation



Search

Embedding Model

Question

"How do I turn off the automatic reverse braking on the Volvo XC60?"

Vector Database

Relevant context from company data

Question + Context

Gen AI Model

Reliable Answer

"The driver can choose to deactivate auto-brake with Rear Auto Brake (RAB) and Cross Traffic Alert (CTA). The warning signal can be deactivated separately. Activate or deactivate the auto-brake with this button in parking camera view"

Access to Current and Comprehensive Information

Up-to-Date Responses: Provides current and updated information by retrieving data from external sources.

Comprehensive Knowledge Base: Extends LLMs' knowledge base with a wide range of information from various sources.

Enhanced Accuracy and Relevance

Reduction in AI Hallucinations: Grounds responses in real-world, verifiable data, reducing the generation of incorrect information.

Contextually Relevant Responses: Integrates relevant external data, ensuring accurate and applicable responses.

Increased User Trust and Transparency

Trustworthy Responses: Offers sourced and current information, increasing user trust in AI systems.

Transparency: Includes citations or sources for external data, allowing users to verify information.

Cost and Resource Efficiency

Reduced Need for Continuous Retraining: Supplements LLMs' knowledge with real-time data retrieval, reducing the frequency of retraining.

Efficient Data Utilization: Utilizes existing data sources efficiently, minimizing the need for large datasets.

Versatility and Customization

Adaptability to Various Domains: Can be tailored to specific industries or domains, making it versatile for different applications.

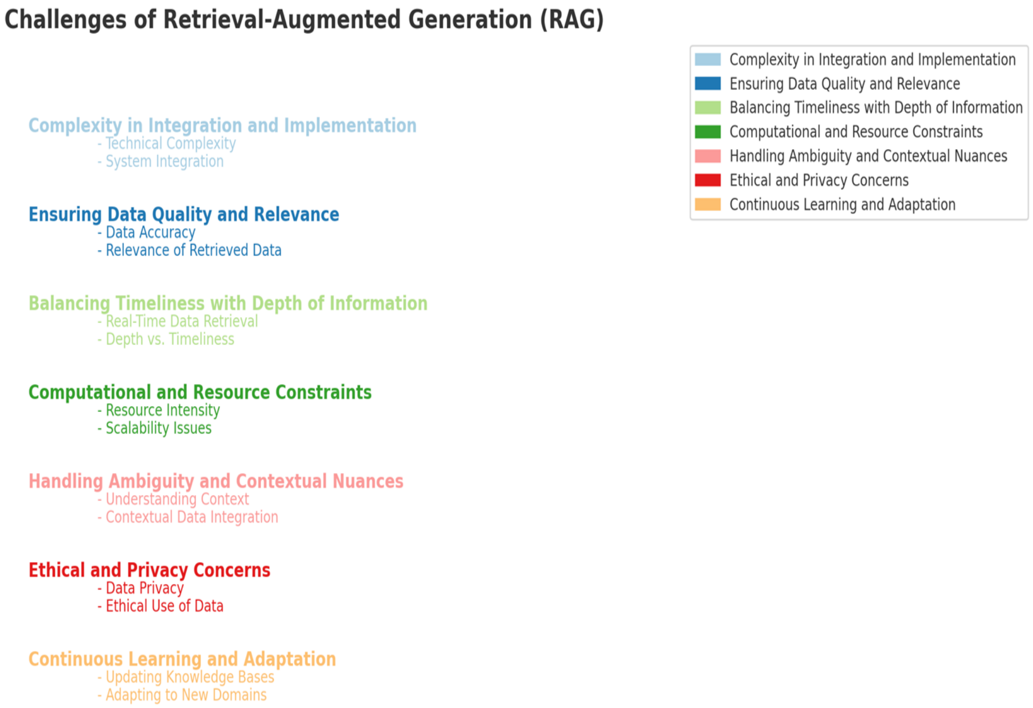
Customizable Responses: Pulls data from domain-specific sources for highly relevant responses.

Improved User Experience

Interactive and Dynamic Interactions: Enhances the interactivity of LLMs, making conversations more engaging.

Meeting User Expectations: Provides accurate, relevant, and current information, leading to better user satisfaction.

Challenges associated with Retrieval-Augmented Generation (RAG)



Challenges of Retrieval-Augmented Generation (RAG)

Complexity in Integration and Implementation
- Technical Complexity
- System Integration

Ensuring Data Quality and Relevance
- Data Accuracy
- Relevance of Retrieved Data

Balancing Timeliness with Depth of Information
- Real-Time Data Retrieval
- Depth vs. Timeliness

Computational and Resource Constraints
- Resource Intensity
- Scalability Issues

Handling Ambiguity and Contextual Nuances
- Understanding Context
- Contextual Data Integration

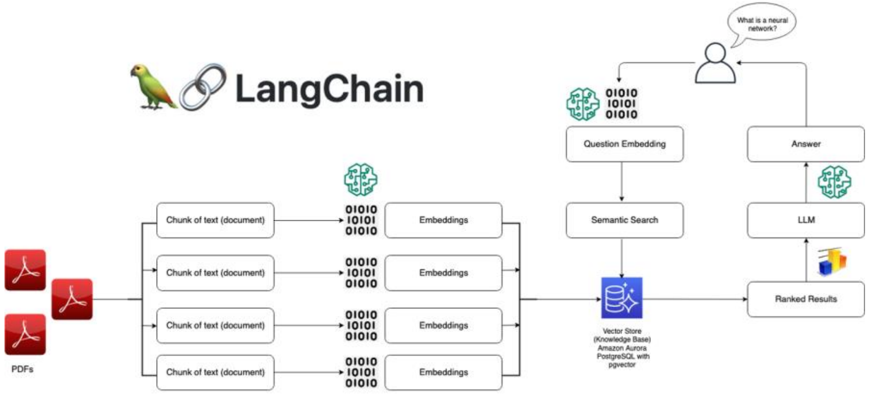
Ethical and Privacy Concerns
- Data Privacy
- Ethical Use of Data

Continuous Learning and Adaptation
- Updating Knowledge Bases
- Adapting to New Domains

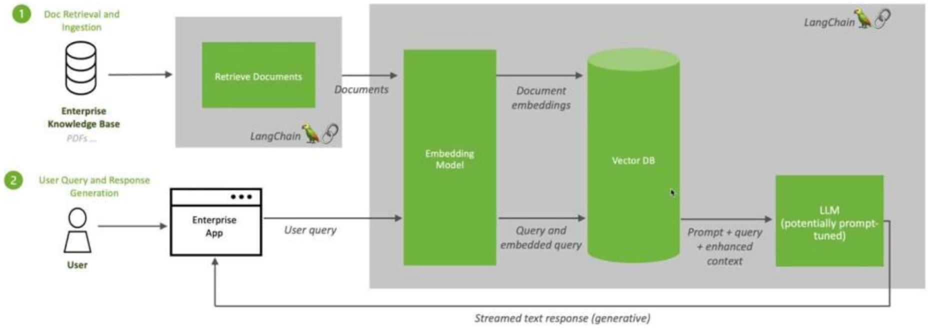
🟦 Complexity in Integration and Implementation
🟦 Ensuring Data Quality and Relevance
🟩 Balancing Timeliness with Depth of Information
🟩 Computational and Resource Constraints
🟥 Handling Ambiguity and Contextual Nuances
🟥 Ethical and Privacy Concerns
🟧 Continuous Learning and Adaptation

Chaining with LangChain

- Chains are sequences of commands, such as LLM calls or data preprocessing steps, created using LangChain Expression Language (LCEL).
- LCEL provides a high-level method for building customized chains.
- LangChain supports LCEL-based chains and legacy chains from previous chain classes.
- Transitioning to LCEL-based chains is motivated by benefits like direct modification of internal components, support for various processing modes, and automatic monitoring at each step.
- This transition enhances flexibility and efficiency in chain creation and execution.

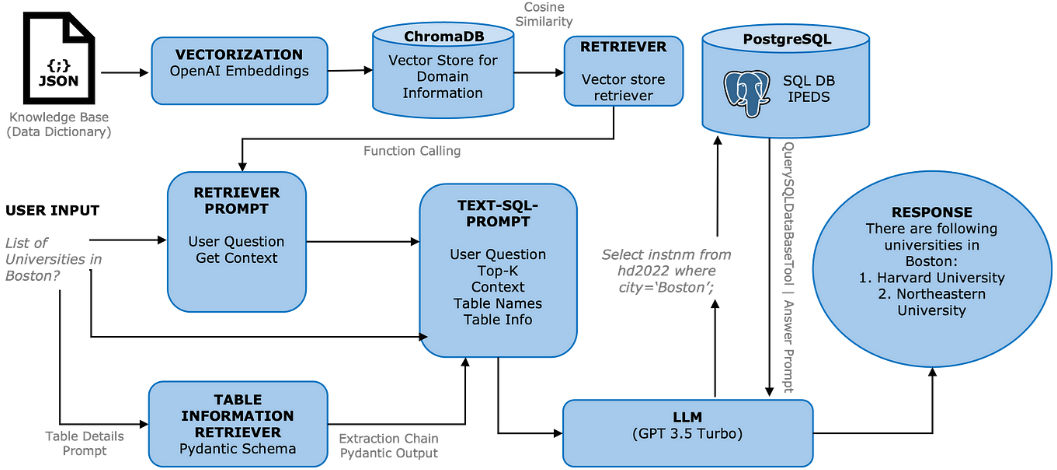


Retrieval Augmented Generation (RAG) Sequence Diagram


1 Doc Retrieval and Ingestion
Enterprise Knowledge Base
PDFs ...
Retrieve Documents
LangChain
Documents
Embedding Model
Document embeddings
Vector DB
LangChain
2 User Query and Response Generation
User
Enterprise App
User query
Query and embedded query
Prompt + query + enhanced context
LLM (potentially prompt-tuned)
Streamed text response (generative)

Context Injection with RAG

- **Establishing a Well-Formatted Knowledge Base**
- Typically in JSON-like format
- Serves as the foundation for creating embeddings and storing them in vector storage (e.g., ChromaDB)
- **Text-to-SQL Workflow with RAG**
- **ChromaDB**
  - User-friendly and open-source
  - Offers supplementary filtering features for metadata association
- **Embedding Function**
  - Uses OpenAIEmbeddings to establish an embedding function
- **Process Steps**
- Load the JSON file
- Apply metadata_func to extract JSON fields for document metadata
- Use content_key to specify the field for vector text (e.g., ‘Table_Name’)
- Instantiate a ChromaDB instance with documents and embedding model
- Construct a chain using the ChatOpenAI model and a retriever



{;} JSON

Knowledge Base
(Data Dictionary)

VECTORIZATION
OpenAI Embeddings

ChromaDB
Vector Store for Domain Information

Cosine Similarity

RETRIEVER
Vector store retriever

PostgreSQL
SQL DB IPEDS

Function Calling

USER INPUT

List of Universities in Boston?

RETRIEVER PROMPT
User Question
Get Context

TEXT-SQL-PROMPT
User Question
Top-K Context
Table Names
Table Info

Select instnm from hd2022 where city=‘Boston’;

QuerySQLDataBaseTool | Answer Prompt

RESPONSE
There are following universities in Boston:
1. Harvard University
2. Northeastern University

Table Details Prompt

TABLE INFORMATION RETRIEVER
Pydantic Schema

Extraction Chain
Pydantic Output

LLM
(GPT 3.5 Turbo)

Off-the-Shelf Chains

- **LangChain Off-the-Shelf Chains**
  - **Chains Created Using LCEL**
    - Built with a more abstract method but essentially constructed using LCEL.
  - **Legacy Chains**
    - Created by subclassing from a previous Chain class.
    - Operate independently from LCEL as distinct classes.
- **'create_sql_query_chain' Using LCEL**
  - **Function Purpose**
    - Produces SQL queries.
  - **Required Parameters**
    - llm: Language model used (e.g., gpt-3.5-turbo-1106).
    - db: SQLDatabase for query generation (a wrapper around SQL Database connection using SQLAlchemy Core API).
    - Text-to-SQL prompt: Prompt used (default is None, chosen based on dialect if unspecified).
    - Top_k: Number of results per select statement (default is 5).
  - **Additional Parameters**
    - top_k: Defines the number of results per select statement, corresponding to the ‘k’ argument.
    - table_info: Receives table definitions and sample rows.
    - table_names_to_use: If specified, only those tables are included; otherwise, all tables are included.
    - dialect: If present in the prompt, the database dialect is conveyed through this parameter.
  - **Returns**
    - A chain capable of receiving a question and generating a SQL query to address it.

```
def get_table_details():
# Read the CSV file containing Table Names and Descriptions using Pandas DataFrame
table_description = pd.read_csv("Data/table_descriptions.csv")

# Retrieving Table Names and Descriptions from the DataFrame
table_details = ""
for index, row in table_description.iterrows():
table_details = table_details + "Table Name:" + \
row['Table'] + "\n" + "Table Description:" + \
row['Description'] + "\n\n"

return table_details

#Creating a Pydantic Base Model
class Table(BaseModel):
"""Table in SQL database."""

name: str = Field(description="Name of table in SQL database.")


def get_tables(tables: List[Table]) -> List[str]:
tables = [table.name for table in tables]
return tables

table_details = get_table_details()
table_details_prompt = f"""Refer the Above Context and Return the names of SQL Tables mentioned
in the above context\n\n
The tables are:

{table_details}
"""

table_chain = {"input": itemgetter("question")} | create_extraction_chain_pydantic(
Table, llm, system_message=table_details_prompt) | get_tables
```

```
# Convert "question" key to the "input" key expected by
current table_chain.
table_chain = {"input": itemgetter("question")} |
table_chain
# Set table_names_to_use using table_chain.
full_chain = (
RunnablePassthrough.assign(table_names_to_use=table_chain) |
generate_sql_query
```

Hands on Tutorial

Colab

https://colab.research.google.com/drive/1_1BAPutSPI1qXmZweFbK2pMP-UlYAq2K?usp=sharing

Hands on Tutorial

SQL Agent

https://drive.google.com/file/d/197eAnyQJC3fwcx4iFa71zJnZmxPbV_-B/view?usp=sharing

Access DB through RAG and LLMs

https://colab.research.google.com/drive/1j-MyhyFlxDRe3Suus4Ev5NhIZui_IMaU?usp=sharing